Comparing temporal behavior of fast objective video quality measures on a large-scale database

Ahmed Aldahdooh*, Enrico Masala†, Glenn Van Wallendael‡ and Marcus Barkowsky*
*LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597, Nantes, France
Email: firstname.lastname@univ-nantes.fr
†Control and Computer Engineering Department, Politecnico di Torino, Torino, Italy — Email: masala@polito.it
‡ Ghent University - iMinds - Data Science Lab, Ghent, Belgium — Email: glenn.vanwallendael@ugent.be

***Abstract*—In many application scenarios, video quality assessment is required to be fast and reasonably accurate. The characterisation of objective algorithms by subjective assessment is well established but limited due to the small number of test samples. Verification using large-scale objectively annotated databases provides a complementary solution. In this contribution, three simple but fast measures are compared regarding their agreement on a large-scale database. In contrast to subjective experiments, not only sequence-wise but also framewise agreement can be analyzed. Insight is gained into the behavior of the measures with respect to 5952 different coding configurations of High Efficiency Video Coding (HEVC). Consistency within a video sequence is analyzed as well as across video sequences. The results show that the occurrence of discrepancies depends mostly on the configured coding structure and the source content. The detailed observations stimulate questions on the combined usage of several video quality measures for encoder optimization.**

***Index Terms*—Video quality, Measure agreement, Large-scale database**

I. Introduction

Typical industrial video distribution chains may continuously monitor the video quality at several processing steps, at the camera capture, on the contribution channel to the studio, for the distribution to the customer, and finally at the customer side. In this work, the application focus would be on those parts where a reference video is available for comparison to a degraded video using Full-Reference (FR) video quality measures and measurement needs to be performed in realtime, potentially on low-performance network equipment. The reference video may either be available explicitly, for example as input to an encoder step, or implicitly, for example using a (camouflaged) test video during regular operation.

A huge number of FR algorithms have been developed and are still in development by researchers in industry and academia ranging from very low to very high computational demands. The evaluation of these methods is usually performed by comparing their prediction performance to ground truth data obtained in subjective experiments, a typical example being the validation experiments by the Video Quality Experts Group (VQEG)[1] that led to several Recommendations of the International Telecommunication Union (ITU-T J.144, J.247, J.341). Performance evaluation by subjective experiments may be seen as mandatory and thus necessary but not sufficient due to the limited number of test cases with respect to the abovementioned application scenario.

Automatic performance analysis only using objective measurements provides a complementary approach. Two alternatives shall be mentioned here. First, the creation of dedicated test sequences in which the performance is expected to be known a priori such as increasing strength of a distortion[2]. The second possibility is to create and evaluate successively a large-scale database. This approach considers that it is not feasible to test the whole database subjectively because it contains an infinite number of potential video sequences. A potential alternative to exhaustive subjective assessment is to compare the agreement of a set of objective algorithms, potentially followed by subjective evaluation of a subset.

Analysis methods and preliminary conclusions using such agreement analysis were proposed by the authors for coding and packet-loss scenarios [3], [4]. The analysis used either pairwise comparisons or additional indicators that were fitted either to improve coherence or to analyze the behavior of the measures. Using the same type of analysis, this paper proposes an evaluation of objective measurements that is difficult to achieve in subjective assessment: Characterization of single frame prediction performance in the context of a video sequence. By framewise analysis, important insight may be gained concerning the scope of application of a measure, for example regarding suitable temporal pooling strategies, i.e. required smoothing for outliers or rate-distortion applications. In the latter case, different distortion measures may be considered in order to improve the smoothness of the perceived video quality optimization. Due to the still limited size of the current large-scale database and the selection of the objective measures dictated by the available processing power, this study focuses on presenting innovative analysis methods rather than generalizable results.

The paper starts with a short summary of the database and the objective measures in Section II. Then, two types of analysis are shown as depicted in Fig. 1. The first, pairwise ranking comparison of consecutive frames as a measure of coherence is presented in Section III. The second is introduced in Section IV in which different source videos and coding parameters are framewise compared providing insight into influence of content and coding structure decisions on coherence. A summary is provided in the conclusion Section V.

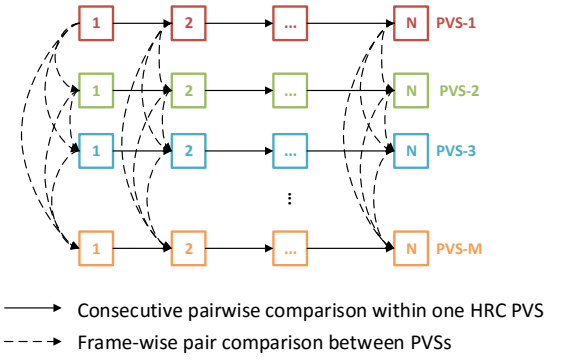

Fig. 1: Illustrations for the two types of analysis that are demonstrated in this paper.

II. Large-scale database description

The HEVC large scale database [3], [5] is created from 10 source videos of 10 seconds long with a wide variation including a cartoon, sports content, nature, and user generated content. The original High Definition (1920x1080) sources have also been downscaled to 1280x720 and 960x544 before further processing by a Hypothetical Reference Circuit (HRC).

As HRCs, only compression has been considered using a varied set of parameters. First of all, the bitrate has been fixed using two constant bitrate techniques (frame based and coding unit based at 0.5, 1, 2, 4, 8, and 16 Mbps) and quantization parameter (QP) based (at QPs of 26, 32, 38, 46). Second, the Group Of Pictures (GOP) size has been varied between two (IBPBPBPBP) and eight (IBBBBBBBP) with one low delay variation having a GOP size of four. Both open-GOP and closed-GOP structures have been considered at intra periods of 8, 16, 32, and 64. Finally, the number of slices has been varied (one, two, and four slices per picture) including a fixed slice size version providing 1500 bytes per slice. In total, 59520 sequences have been produced in this way enabling a data analysis approach on video compression behavior.

In this work, the strategy has been to start with a limited set of sources and a large variety of compression parameters or HRCs in order to keep processing feasible. In a later phase, by identifying the most useful subset of HRCs an extension of the number of sources is planned against this restricted set of HRCs. From all these encoded sequences, i.e., Processed Video Sequences (PVS), the frame based and sequence average of Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [6], and Visual Information Fidelity (VIFP) [7] have been calculated.

III. Consistency measure on consecutive frames

In this section, the continuity of agreements and disagreements is analyzed within one HRC, i.e. within one coding condition, for each source content.

$$Agreement = \begin{cases} 1 & |\sum_{Q\in\{PSNR,SSIM,VIFP\}} sign(Q(A)-Q(B))| = 3 \\ 0 & else \end{cases} \quad (1)$$

The continuity of agreements is measured in a sequential manner as formally described in Eq. 1 and as illustrated in Fig. 1 with solid arrows. Once the disagreement is happening between frames (A and B), there is high probability that frame B disagrees too with frames before A. For the given large-scale database, there are 5952 HRCs for each 250-frame source. Hence, a 5952x249 agreement/disagreement matrix is calculated for each source. Then, for each of these sources, the columns are summed and divided by the number of HRCs. This type of analysis shows the temporal behavior of different objective video quality metrics, namely PSNR, SSIM, and VIFP. Fig. 2 shows the variations of the number of disagreements over time for two source contents: source number 6 and 10. The darker the bar for a particular frame, the higher the fraction of disagreement between the frame represented on the X-axis and its previous frame. It is difficult to make general interpretations of the maxima from such an overall analysis. A better strategy is to consider agreement with respect to the different sources or coding conditions as described in the following subsections as well as in Section IV. On the other hand, what can already be observed in this high-level analysis is that the highest number of agreement (white peaks) are happening when agreement between video quality measures is calculated between Intra-frames and their successive/preceding Inter-frame. Further analysis of the data reveals that this is because the Intra-frame has a notable higher quality compared to next/previous Inter-frame from the quality measure point of view. The used encoder configurations implies a higher quality to the Intra-frame compared to Inter-frames such that all measures easily agree on which frame is highest in quality. Thus, when measuring improvements of newly proposed algorithm, it is advantageous to compare all available objective measures with respect to the content in order to provide a thorough analysis of the proposal.

A. The impact of content

When analyzing the data in more detail, the influence of the content types and characteristics clearly appears. From the data, it can be observed that the number of disagreement varies from one content type to the other. In Fig. 3, the fraction of disagreement for each quality measure is displayed. It can be observed that the contribution of each quality measure to the overall disagreement is very clear. The majority of disagreement in SRC3 is due to PSNR, while the majority of disagreements in SRC10 is due to SSIM, the figure is not presented here due to space limitations. From these observations, it can be concluded that depending on the type of the source content, PSNR, SSIM, and VIF can act differently. Thus, when measuring improvements of algorithms, it is advantageous to compare all available objective measures with respect to the content in order to provide a thorough analysis.

B. The impact of Intraperiod

As mentioned in the high-level analysis, the Intraperiod is a very important factor in understanding the temporal behavior of the quality measures. Fig. 4 shows this effect. It demonstrates the variation of disagreements of HRCs of SRC6.

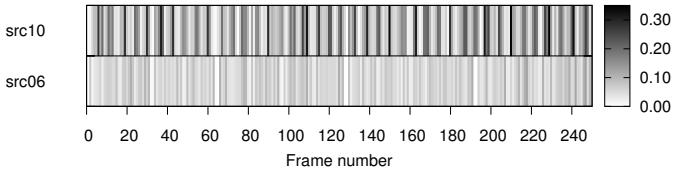

Fig. 2: The variations of the number of disagreements over time for two source contents: source number 6 and 10.

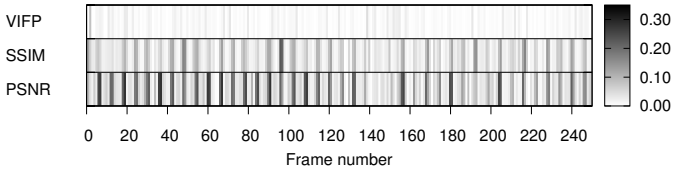

Fig. 3: The cause of disagreements in SRC3.

It is obvious that the disagreement fractions between Intra-frames and next/previous frames is very low compared to other frames. Similar observations can be made for all Intraperiods (8, 16, 32, and 64) and also for the other source contents. The capability of the quality measures to agree when comparing two frames of notable difference in quality is the main reason for this phenomenon. Hence, when a source is encoded with coding conditions that only differ in the Intraperiod, temporal pooling strategies for calculating the video quality score may be examined and this effect may be taken further into account.

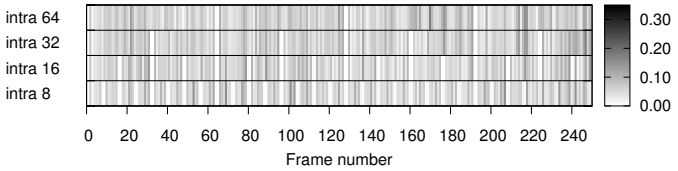

Fig. 4: The variation of disagreement fractions of HRCs with Intraperiod of 8, 16, 32, and 64 of SRC6.

C. The impact of GOP structure

Low-delay and the hierarchical structure of GOP configurations are widely used in different application scenarios. In this work, the consistency of quality measures is categorized to show the role of hierarchical GOP structure with different sizes and a low-delay configuration of size four. Fig. 5 shows this role for SRC6. The number of disagreement in the low-delay configuration is higher than the number in the hierarchical coding structures. This observation stands for all source contents except for SRC3. This behavior of low-delay might be due to the its configuration of using not only the previous frame but also -5,-9, and -13 frames relative to the first frame of the GOP. Moreover, in low-delay there is only one layer and the quality of the inter-frames are very similar, which yields a high inconsistency between the quality measurements.

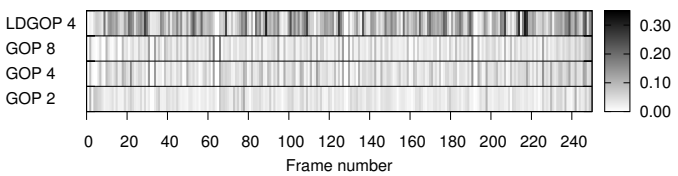

Fig. 5: The variation of disagreement fractions of HRCs with hierarchical GOP of 2, 4, and 8 and the low delay of 4.

D. The impact of QP and rate control

An interesting observation can also be made for the impact of using a constant quantization parameter or a rate control configuration. Very low and very high disagreement fractions periodically alternate at the beginning and the middle of the GOP while this is not observed when rate control is used. Fig. 6 shows this observation for SRC7. In this source, the fraction of disagreements for some frames is higher than 50% when constant QP is used while it is not the case for the rate control option.

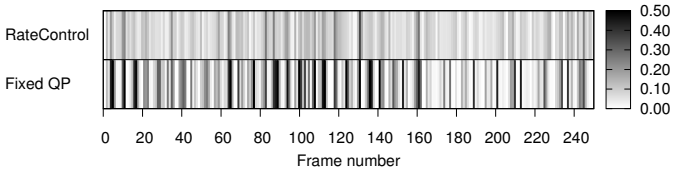

Fig. 6: The variation of disagreement fractions of HRCs with constant QP and rate control option.

IV. Consistency with respect to source content and coding parameters

In this section the agreement of the measures is analyzed across PVS, i.e. by considering two PVS at a time and comparing their quality measure values for each single frame (see the dashed arrows in Fig. 1). Consider, for instance, the sequence-level values of each of those metrics for two different PVSs. Two cases are possible: either all the measures agree (Case *Agree*) on which PVS provides the best quality, or they do not agree (Case *Disagree*). From this point, we only consider Case *Agree*, and we investigate if such an agreement at the overall sequence level corresponds to agreement for single frames as well.

First, we observed that, for sequences for which the quality is strongly different, typically there is agreement at the frame level, i.e., comparing the measures for frames in the same position in the two sequences yields to agreement among the measures. However, when the quality difference is less pronounced, even in Case *Agree*, for some frames in the sequence there is no agreement for frames in the same position. For the purpose of this work we consider only sequences for which the agreement holds for more than 90% of the frames (Case *Agree90*). The rationale behind this choice is that when a new coding and/or processing technique is proposed, typically quality values for the overall sequence are presented to show that the new technique is better than some reference. In absence of further information, such form of presentation typically creates the expectation that the improvement holds for the large majority of the frames in the sequence. If this is not the case, it might be a symptom of some temporal irregularities that should be better investigated directly by the proponents.

In the rest of the section, we will focus on Case *Agree90* by investigating how the disagreement between corresponding frames in different sequences is influenced by the coding parameters. By fixing the value of most of the coding parameters described in Section II, we obtain a set of sequences from

which we choose the Case *Agree90* ones. The latter ones are compared one against each other, yielding to $N(N-1)/2$ comparisons when N sequences are considered.

As a first example, we consider the number of slices per frame. Fig. 7 shows, for each frame position, the fraction of frames in that position that disagree among all the performed comparisons, and for which the reason of disagreement is the PSNR. This operation is repeated for similar sets in which only the number of slices per picture changes. It can be observed that the number of frames and their temporal position is very similar, therefore it seems that the number of slices does not significantly affect the number of disagreement.

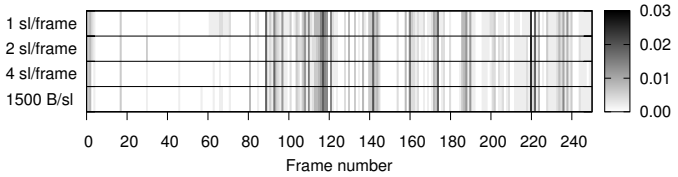

Fig. 7: Fraction of frames in disagreement for different number of slices per frame. Fixed QP, GOP size 8, intra refresh 16, open GOP.

This method allows to intuitively see the difference and their position for a few different conditions, however it is impractical to perform large scale analysis. Therefore, instead of visually comparing the behavior over time of the fraction of disagreement, we propose to compute a similarity index, i.e., the absolute value of the correlation coefficient. Such an approach also allows to provide a quantitative measurement of the similarity.

	1 sl/frame	2 sl/frame	4 sl/frame	1500 B/slice
1 sl/frame	1.000	0.988	0.976	0.969
2 sl/frame	0.988	1.000	0.974	0.973
4 sl/frame	0.976	0.974	1.000	0.966
1500 B/slice	0.969	0.973	0.966	1.000

TABLE I: Correlation coefficient among the results of Fig. 7.

The previous figure can be compactly represented by the data in Table I. To further improve the scalability of the method, we represent such data using matrices with different gray values, where the darker is the gray level, the higher is the absolute correlation. Fig. 8 shows the same data of the previous table in this form. The image is obviously symmetric along the diagonal as the values in the table.

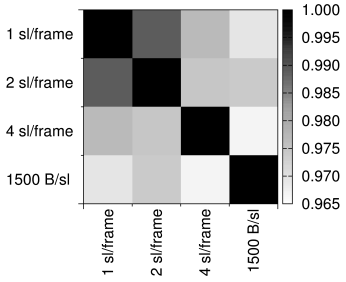

Fig. 8: Graphical representation of correlation coefficients shown in Table I.

We adopt this technique to analyze the influence of the major coding parameters. When the correlation is close to one, the parameter has almost no impact, whereas lower values show much higher influence.

First, we consider the fixed quantization parameter (QP) case, as done in most of the video coding works [8], and we vary only one parameter at a time. When all combinations of all the other coding parameters, including the source sequence, are considered, instead of only a subset as in Fig. 8, results are similar, as shown in the left part of Fig. 9.

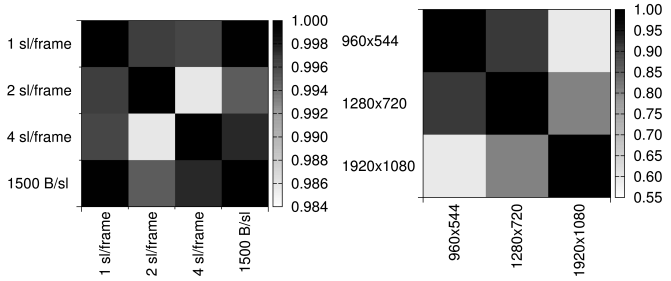

Fig. 9: Correlation coefficients between the cases in which all but the slice size parameter (left) and resolution (right) are varied.

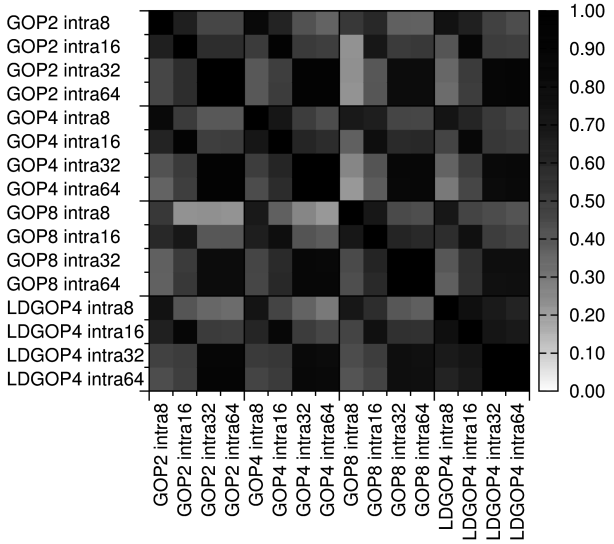

Fig. 10: Correlation coefficients between the cases in which all but the GOP size and intra refresh parameters are varied.

The same behavior happens for the open or closed GOP parameter (not shown in figures, correlation equal to 0.906), and partly for the resolution as in the right side of Fig. 9.

The more interesting parameters are the Intraperiod and the GOP size. Significant variations can be observed, especially when they are considered jointly as in Fig. 10. In particular, it seems that when the GOP size is small and the intra period is large, there might be a strong impact on the position of disagreements, whereas with the largest GOP size the effect is reduced. With the low-delay GOP configuration (LDGOP) correlation is very high, meaning that the influence of the intra refresh rate is much more reduced.

When the rate control algorithm of the HM test model software [9] is used instead of the fixed QP parameter, interesting observations can be made in the data, in particular when they are represented as a function of the frame position in the sequence. Fig. 11 is an example of such condition. The two rows are almost equal since they only differ for the open

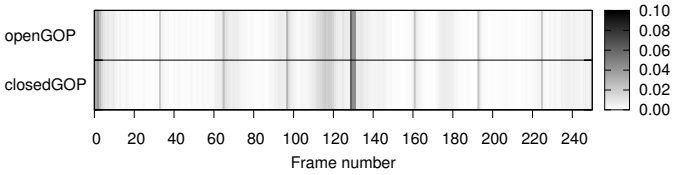

Fig. 11: Fraction of frames in disagreement for different number of slices per frame. HM rate control, LDGOP size 4, intra refresh 32. Note the peaks (darker vertical lines) at multiple of 32 frames.

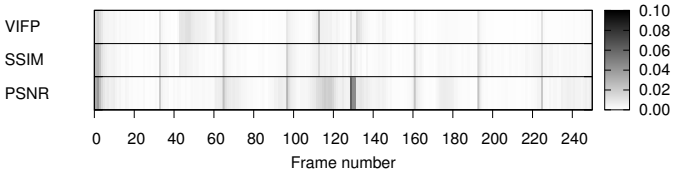

Fig. 12: Fraction of frames in disagreement separated for each measure. HM rate control, LDGOP size 4, intra refresh 32, open GOP.

or closed GOP parameter which, as previously stated, has very little influence.

For instance, in the first part a high fraction of disagreement is visible. This can be ascribed to the fact that an initial, fixed, QP is used by the HM rate control algorithm, which then quickly adapts to the requested bitrate.

Moreover, note the peaks which appear in correspondence of the periodicity dictated by the intra refresh rate, i.e., when frames with I-type blocks only are inserted. By further experiments we determined that this behavior is probably due to the inclusion of some source sequences which seems particularly difficult for the HM rate control when a frame with I-type blocks is inserted.

This observation underlines the importance of performing such types of analysis on a large database with multiple coding parameters and several different content types. Although our database is somehow limited in the latter aspect, nevertheless such effects can already be observed.

Finally, we consider the fraction of disagreement by considering the same sets of comparisons but computing the fraction of disagreement for all the three measures. Figure 12 shows an example of the typical situation. While some behaviors are common for all metrics, e.g., the initial frames and the periodicity of the peaks, others seem to be peculiar of the measures. However, the latter often have a lower intensity.

V. Conclusions

It may have been expected that disagreement between several objective measures exists on a frame-level even if the measures agree on a sequence level. However, the particular patterns of this disagreement point to two important conclusions. The first conclusion is that the usage of one single measure may not be sufficient. In particular, it may be beneficial to analyze the usage of several complementary algorithms within the coding loop, i.e. for rate-distortion optimization. In addition, it should be noted that performance bias may occur when improvements are measured only objectively and only using one single method, thus weakening such proposals. The second conclusion is that the pronounced correlation between content characteristics and encoder parameter selection encourages further analysis, for example with respect to the efficiency of rate-control algorithms. Some coding factors are almost not influential, whereas others have a strong impact, suggesting that quality comparisons among sequences without considering the detailed behavior of the quality over the frames in the sequence itself could be strongly misleading. While these results concern future developments of coding and quality measurement algorithms, further work on the large-scale database approach requires a significant extension of the samples, both sequences and algorithm results, which is currently limited by the computational resources and the availability of implementations of objective measurement algorithms. Methodical work on analysis methods using statistical methods will continue towards the identification of particular cases that require inclusion in subjective experiments and the characterization of objective measures.

Acknowledgment

Some of the computational resources have been provided by HPC@POLITO (http://www.hpc.polito.it). Some parts of this work are supported by the Marie Skłodowska-Curie under the PROVISION (PeRceptually Optimised VIdeo CompresSION) project bearing Grant Number 608231 and Call Identifier: FP7-PEOPLE-2013-ITN. The research activities described in this paper were partially funded by Ghent University, iMinds, Flanders Innovation & Entrepreneurship (VLAIO), the Fund for Scientific Research Flanders (FWO-Flanders), and the European Union. Some aspects of this work were carried out using the STEVIN Supercomputer Infrastructure at Ghent University.

References

[1] "Video Quality Experts Group (VQEG)," Jul. 2016. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/vqeg-home.aspx

[2] H. Liu and A. R. Reibman, "Software to stress test image quality estimators," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, June 2016, pp. 1–6.

[3] G. Van Wallendael, N. Staelens, E. Masala, and M. Barkowsky, "Full-HD HEVC-encoded video quality assessment database," in *Ninth International Workshop on Video Processing and Quality Metrics (VPQM)*, 2015.

[4] A. Aldahdooh, E. Masala, O. Janssens, G. Van Wallendael, and M. Barkowsky, "Comparing simple video quality measures for loss-impaired video sequences on a large-scale database," in *Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, Jun. 2016, pp. 1–6.

[5] M. Barkowsky, E. Masala, G. Van Wallendael, K. Brunnstrom, N. Staelens, and P. Le Callet, "Objective video quality assessment – towards large scale video database enhanced model development," *IEICE Transactions on Communications*, vol. E98-B, no. 1, pp. 2–11, Jan. 2015.

[6] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[7] H.R. Sheikh and A.C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[8] F. Bossen, "Common test conditions and software reference configurations," *Doc. JCTVC-J1100*, Jul. 2012.

[9] K. McCann, B. Bross, W.-J. Han, I.-K. Kim, K. Sugimoto, and G. J. Sullivan, "High Efficiency Video Coding (HEVC) Test Model 12 (HM 12) Encoder Description v. 12.1 Doc. JCTVC-N1002," Nov. 2013.